Procedia Computer Science

Volume 29, 2014, Pages 379–389

ICCS 2014. 14th International Conference on Computational Science

Spatio-temporal Sequential Pattern Mining for Tourism Sciences

Luke Bermingham and Ickjai Lee

School of Business (IT), James Cook University, Cairns, Queensland, Australia
luke.bermingham@my.jcu.edu.au, Ickjai.Lee@jcu.edu.au

Abstract

Flickr presents an abundance of geotagged photos for data mining. Particularly, we propose the concept of extracting spatio-temporal meta data from Flickr photos, combining a collection of such photos together results in a spatio-temporal entity movement trail, a *trajectory* describing an individual's movements. Using these spatio-temporal Flickr photographer trajectories we aim to extract valuable tourist information about where people are going, what time they are going there, and where they are likely to go next. In order to achieve this goal we present our novel spatio-temporal trajectory regions-of-interest mining and sequential pattern mining framework. It is different from previous work since it forms regions-of-interest taking into consideration both space and time simultaneously, and thus produces higher-quality sequential patterns. We test our framework's ability to uncover interesting patterns for the tourism sciences industry by performing experiments using a large dataset of Queensland photo taker movements for the year 2012. Experimental results validate the usefulness of our approach at finding new, information rich spatio-temporal tourist patterns from this dataset, especially in comparison with the 2D approaches shown in the literature.

Keywords: Sequential pattern mining, Spatio-temporal, Tourism science, Movement pattern

1 Introduction

Data mining is the extraction of interesting, previously unknown knowledge from potentially large and noisy datasets. We reason that data mining could be especially valuable to the field of tourism science where abundant amounts of information regarding people's movements and activities is available, yet untapped. Specifically, we refer to advances in camera and mobile phone technology, these devices are now able to record spatial and temporal (spatio-temporal) meta data every time a photograph is taken. Combining a collection of photographs together we can formulate the approximate movement trails of an individual, we call this a *trajectory*. Coupling this effect with the advent and popularity of photo sharing sites such as Flickr we propose that there now exists a unique opportunity to extract previously unknown, valuable patterns from a massive amount of spatio-temporal photo-taker trajectories. Furthermore, when we consider

Selection and peer-review under responsibility of the Scientific Programme Committee of ICCS 2014

doi:10.1016/j.procs.2014.05.034

that many photos are being taken in known tourist locations and the fact that they are being shared through social media we can infer that some of these photo-takers are tourists, thus leading to the valuable conclusion: we can know where tourists are moving and at what times they going there. This information is clearly valuable to the tourism sciences industry and as such is a strong motivation for our study.

Previous research in this area, notably [3], introduces the notion of extracting Regions-of-Interest (RoIs) from trajectory data and then performing Sequential Pattern Mining (SPM) to find popular movement paths between these RoIs. However, no previous work focuses especially on tourism information and additionally no approach can fluidly discover and visualise results in three dimensions. For example in [2] only the spatial dimension is considered during RoI extraction and visualisation. Hence current approaches cannot extract valuable spatio-temporal patterns like which months tourist locations are popular to visit. Therefore, in this research we introduce our truly spatio-temporal trajectory RoI and SPM framework (see Section 3), in order to validate our framework we attempt to find new patterns using the same Flickr data that was used in [1] (see Section 4), from our tests we summarise our findings, draw our conclusions about the validity and effectiveness of our approach, and finally introduce some future directions (see Section 5).

2 Literature Review

2.1 RoI Mining

Trajectory RoI mining was first introduced by Giannotti et al [3]. The notion Giannotti et al propose is that areas of density in the study region are interesting. They propose partitioning the study region into rectangular grid cells and then calculating that number of trajectories that pass through each cell. This grid of dense cells is traversed in order of highest density cells first. Dense cells are expanded rectangularly to form RoIs. Cai et al [4] argue that rectangular expansion produces RoIs that contain uninteresting low-density cells, and therefore they present and validate their novel arbitrary RoI formulation methods: "Slope" and "Hybrid". Additionally, Cai et al [1, 2] test the original Giannotti approach and their approaches on Flickr datasets, showing that the extraction of interesting spatial patterns is merited. Clearly, RoI mining is of interest in extracting knowledge from trajectory datasets. However none of the methods in the literature considers the temporal dimension while mining and thus cannot uncover the valuable spatio-temporal regions that are characteristic of trajectory datasets, thus Flickr photo-taker movements require the additional dimensional partitioning to uncover more meaningful patterns.

2.2 Trajectory SPM

The notion of trajectory SPM, similar to trajectory RoI mining, is also first introduced by Giannotti et al [3]. They take the RoIs extracted from the data, label them, translate the trajectory dataset into a sequence of RoI visitations and then perform the generic SPM algorithm PrefixSpan [8] to discover frequent visitation sequences between RoIs. Due to the fact that the RoI mining [3] only considers the spatial dimension in an attempt to uncover temporal information regarding its sequential patterns it makes the inclusion of a temporal parameter that determines whether patterns are occurring within some specified time range. The problem with this approach is that the results do not necessarily discover RoIs that grow

in the temporal dimension. Performing SPM in such a way does not equate to discovering true spatio-temporal movements between RoIs in space-time, rather it discovers spatial movements that fall into an acceptable time tolerance of each other. The goal of SPM is simply to extract frequent sequences, thus for both computation sake and pattern quality consideration of the temporal dimension should occur before the SPM phase, not during. Further research also performs trajectory SPM using the same spatial and then temporal approach [1, 2, 4], however we reason that spatio-temporal RoI mining and then SPM will produce more comprehensible, high-quality, and previously unobtainable results.

2.3 Spatio-temporal Tourist Patterns

Discovering tourist patterns from Flickr or movement datasets is not a new pursuit. Previously research has been conducted on detecting popular landmarks and travel sequences [5, 6, 7, 9]. However, no approach considers finding true spatio-temporal tourist patterns from Flickr trajectory data. Spatio-temporal tourist patterns are arguably far more valuable than the trip planning or tourist hot-spot detection of other research because it not only describes where the tourist are but also what time they will be there, and then where they are likely to go next. Such information is highly valuable to the industry of tourism science in terms of marketing, pricing, and packaging. Therefore, our concept is to perform spatio-temporal RoI mining on specific Flickr datasets to discover insightful regions that tourists visit at a particular time and then perform SPM to determine if there is a relation between popular spatio-temporal RoIs, i.e people going from one place to another at a particular time of the year.

3 Spatio-temporal RoI and SPM Framework

Our spatio-temporal framework is an extension of that introduced by Giannotti et al [3] and includes the modified RoI mining methods by Cai et al [4]. The SPM mining phase is completely the same except for the omission of the temporal time range criteria which is no longer necessary, meaning sequential patterns are still extracted using PrefixSpan [8]. The details of our framework for discovering spatio-temporal tourist patterns from Flickr data is outlined in more detail in Figure 1 and the accompanying explanation of each component.

1. **Flickr Geotagged Photos.** A huge source of geotagged photos exist online in the form of the photo-sharing website Flickr.

2. **Extract and Preprocess ST Data.** A large collection of geotagged photos are collected using the Flickr API [1], all within our particular study region and time range, for example Queensland in the year of 2012. The specifics of photo information collection process we use are the same as those presented in the methodology by [1]. The preprocessing we apply is simply to remove erroneous entries.

3. **ST Trajectory DB.** Once we have extracted the geotagged photos from Flickr we store them in a spatio-temporal database for fast retrieval and querying. Specifically we ensure each photo-taker has their own trajectory movement trail in the database, and that it is ordered chronologically.

4. **RoI Mining.** We divide the study region into spatio-temporal grid cells, then due to its effectiveness we apply the same "Hybrid" RoI expansion method as in [2]. The only extra

[1] http://www.flickr.com/services/api/

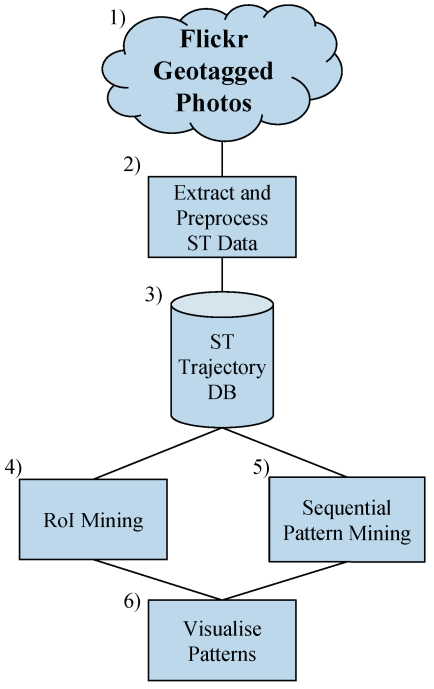

Figure 1: Spatio-temporal RoI mining and SPM framework.

modification we have to make is that spatio-temporal grid cell expansion has to occur in two extra directions now, specifically we expand RoIs {left, right, up, down, front, back}.

5. **Sequential Pattern Mining.** We translate trajectories into RoI visitation sequences and then apply the original implementation of PrefixSpan as our SPM method.

6. **Visualise Patterns.** The visualisation can occur either for RoI mining, because RoIs alone describe interesting spatio-temporal patterns themselves or it can occur for the sequential patterns mined. Due to the visualisation being in 3D the results are more comprehensible than the flat approaches shown in the literature and it is simple to understand where and when patterns are occurring. The actual visualisation is implemented in WorldWind SDK [2].

4 Application to Flickr Data

In this section we present our experimentation on Flickr photo taker data. We attempt to find novel spatio-temporal RoIs and sequential patterns from the dataset and then offer a comparison to the traditional 2D approaches in terms of both effectiveness and efficiency.

[2] http://worldwind.arc.nasa.gov/java/

4.1 Queensland Flickr Dataset

The dataset we choose to use for experimentation is a collection of geotagged photos over the whole of Queensland for the year 2012. The dataset contains 2,445 trajectories in total, made up of 63,290 of photo taker spatio-temporal entries. A visual overview of the dataset is shown in Figure 2. This dataset is highly busy, photo-takers can go from one place to another not necessarily in a straight line, there is no rules governing recording times, and in general there is no underlying structure. In general the data falls along the populated areas of Queensland, the east coast in particular, which is desirable because known tourist locations lie in these regions. In terms of the temporal aspect there exists no human identifiable trend for the whole datasets that describes at what time of the year photos are being taken. Overall, we reason that this a good candidate to test our framework to see if it can uncover underlying patterns in difficult circumstances. The other contributing factor for choosing this dataset is that it is used in [1], and thus a comparison between 2D and 3D RoIs and sequential patterns can be made.

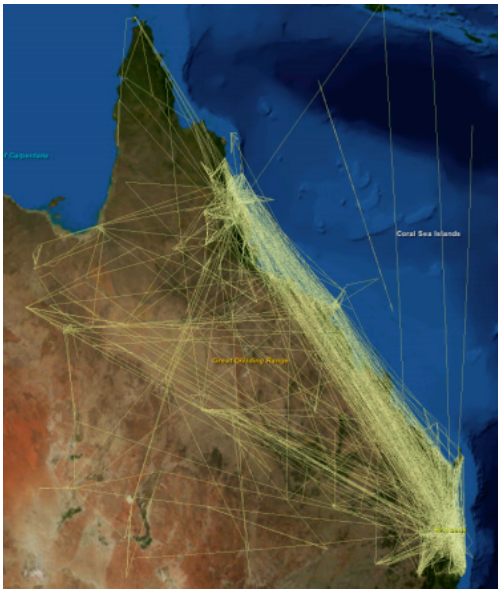

Figure 2: Queensland 2012 Flickr photo taker trajectories.

4.2 Spatio-temporal RoIs

Applying our approach to the Queensland data we are able to identify a number of spatio-temporal RoIs along the east coast. These spatio-temporal RoIs are shown in Figure 3. Figure 3 uncovers interesting patterns emerging in the Queensland data in specific temporal windows. In this experiment the temporal cell size was a fortnight and the spatial cell size 10km. This results shown that towards the end of the year Cairns, Townsville, and Sunchine Coast become popular photo taker destinations, clearly reflecting the seasonal nature of tourism in these cities. Whereas in contrast the framework reports that Brisbane and Gold Coast experience strong photo taker flow all year round, which if we consider this in terms of tourist data makes sense, as they are major hubs in Queensland and have many more year round attractions. For comparison sake, in Figure 4 we include the RoIs discovered by a spatial only query run with

looser parameters. The results shown more RoIs, though the major RoIs found in our spatio-temporal approach remain, thus validating our approach is at least as effective as traditional approaches. Furthermore, through visualisation is it clear to see our spatio-temporal RoIs are more meaningful, they easily display not only where interesting places are, but when people generally visit them too. Using traditional approaches it would be impossible to find seasonal visitation patterns such that we have found, thus validating the merit of our approach for tourism science.

Figure 3: Queensland east coast spatio-temporal RoIs.

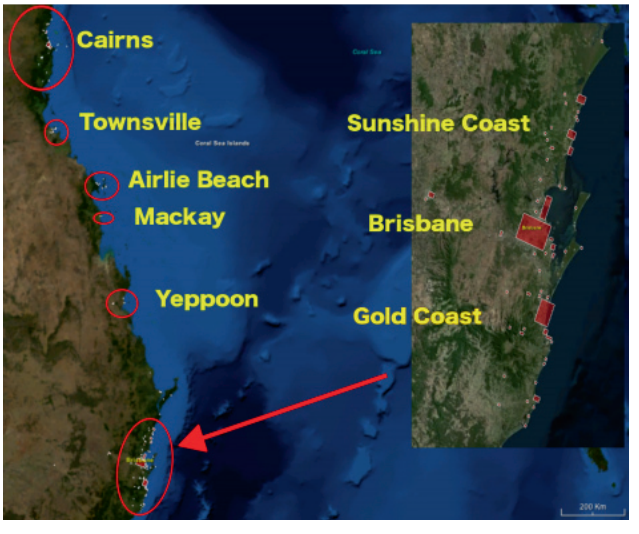

Figure 4: Queensland east coast 2D RoIs, source from [1].

Investigation into the dense seasonal patterns in Brisbane reveals further valuable spatio-temporal knowledge. In Figure 5 we present our findings using cell density parameters that are one standard deviation away from the average cell density. Specifically in a) we show the

monthly dense regions of Brisbane and surrounding area, noting that areas such as Brisbane CBD, SouthBank, and GoldCoast form RoIs all year round whilst other area such as Sunshine Coast only garner attention from the middle of the year onwards. We propose areas that show clear seasonal patterns will correlate strongly with popular tourist visiting times. Furthermore, in b) we show the daily RoIs over the whole of Brisbane and wider regional study space. The daily dense candidate cells compound on top of each other to reveal peak weeks when Brisbane and surrounding regions are all together most popular for photo-takers. It is reasoned that these patterns describe popular general times that tourist and holiday activity is peaked in these areas. For example, in b) the RoIs precisely match up with Christmas and June/July holidays. This kind of result is a strong validation for the usefulness and accuracy of our spatio-temporal mining approach.

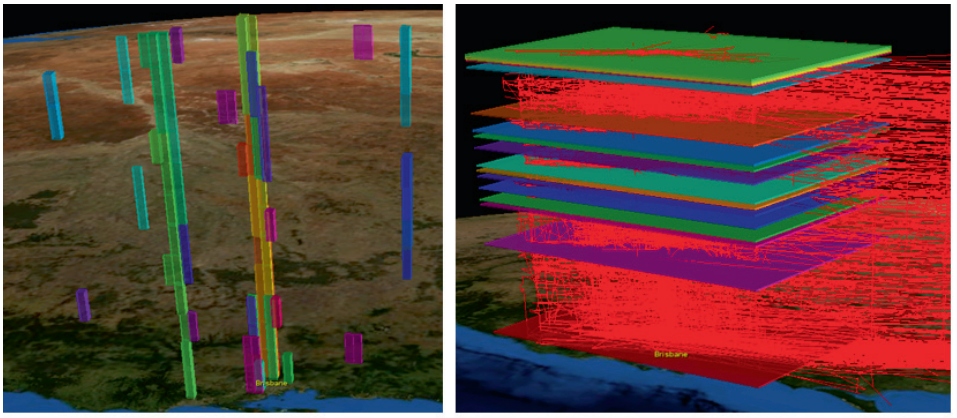

Figure 5: Seasonal patterns in Brisbane: a) 30 day temporal cell size, 5km spatial cell size; b) 24 hour temporal cell size, 300km spatial cell size.

4.3 SPM with Spatio-temporal RoIs

Once spatio-temporal RoIs have been extracted we apply SPM to discover interesting linkages between spatio-temporal hotspots. In Figure 6 we show the results of SPM mining having used spatio-temporal RoI mining and traditional 2D mining. The mining parameters we use for our spatio-temporal mining are a 5km spatial cell size, 30 day temporal cell size, a 1% pattern support, and a cell density of 0.003% (8 trajectories). The first observation we make is that there is a clear division between regions visited when spatio-temporal RoI mining is applied, highlighting its ability to extract more detailed sequential patterns. Overall, we can say the results contain travel sequences from roughly the same regions in the Brisbane area, this validates that our approach still finds major patterns. The main difference is of course the extra information available by including the temporal dimension. We can now see at what times these patterns are occurring. In this case we apply mining with one month temporal window, we can see a linkage: Raby Bay to South Bank to Botanic Garden and Bramble Bay to South Brisbane. We could find these patterns before, but now we can make the interesting observation that there is months of gap between these visits. If we consider this in terms of tourists we can say that tourists who visit the botanic Raby Bay at the start of the year are

likely to visit the botanic gardens later on in the year when the scenery is better. For example we can say that tourists visiting one scenic area have a strong likelihood of visiting another scenic area later on in the season. This is the kind of linkage that could not be observed when only considering the spatial dimension, it is highly relevant to tourism science and thus shows the usefulness of our approach.

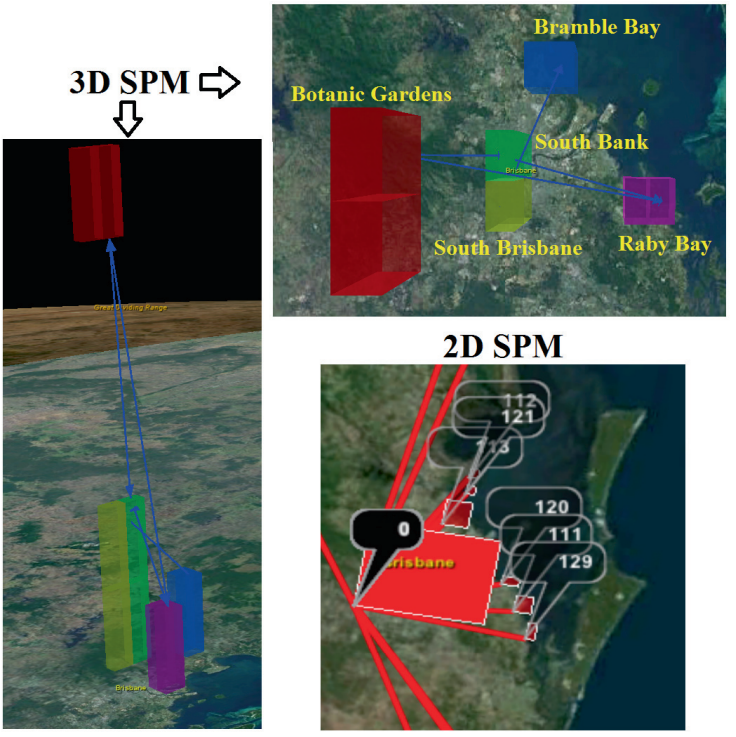

Figure 6: Comparison of SPM 3D and 2D.

Further experimentation into the dense regions surrounding Brisbane reveal more spatio-temporal patterns. We perform 2D and 3D RoI mining in Brisbane and its surrounding area again using cell density that is one standard deviation beyond the average for the study region. The results of the RoI mining are shown in Figure 7. In Figure 7 both a) and b) identify that photo taker movement path from Gold Coast to Brisbane. However in a) we are able to highlight at exactly what time of the year the this transit occurs, namely at the start of the year in summer. Whereas in b) we can detect the pattern but are unable to get such specific and valuable information regarding when it occurs. This opportunity to extract higher quality information sequential patterns by using our 3D RoI mining approach strongly highlights its usefulness and advantage over traditional two dimensional approaches.

4.4 Efficiency of Our Approach

To maintain the usefulness of RoI mining to Flickr trajectory data mining we tested our approach's running time with the following parameters: static temporal cell size of 30 days, a 1% cell density support (which for the Queensland Flickr data equates to a cell density of 24

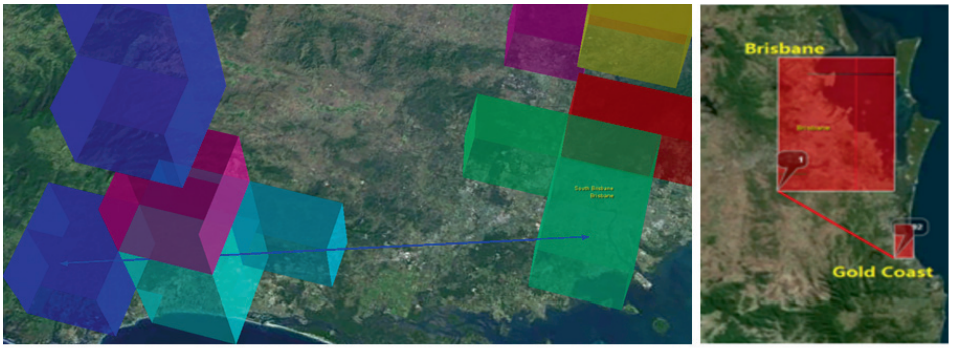

Figure 7: Gold Coast to Brisbane SPM with 10km spatial cell size, 30 days temporal window: a) Spatio-temporal; b) Spatial then Temporal SPM.

trajectories), and a range of spatial cell sizes from coarse to granular. The temporal cell size and the cell density remain the same so that the number of candidate grid cells scales linearly with the spatial cell size. The results of this experiment are illustrated in Figure 8. Figure 8 shows the efficiency of our spatio-temporal RoI mining approach, even with the inclusion of extra dimensionality, density-based RoI mining is performed in milliseconds. However, we do note that by increasing the number of grid cells for mining machine memory is quickly filled, once the memory is full application performance degrades significantly. Specifically in this experiment a spatial cell size of 1km produced such a large number of candidate grid cells that RoI mining was not possible on the 6Gb test machine.

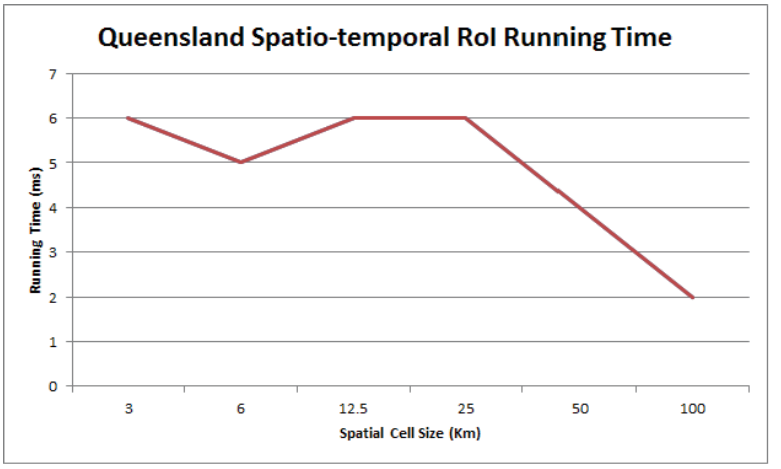

Figure 8: Spatio-temporal running time.

The dual function of our approach is SPM so we must test that it remains affordable too. Figure 9 illustrates the running time of our approach and the 2D SPM methods in the literature. The results show both have an exponential running time, this severely limits the usefulness of mining huge Flickr datasets for tourist information because it will exponentially increase as the

data increases or as the parameters becomes more granular.

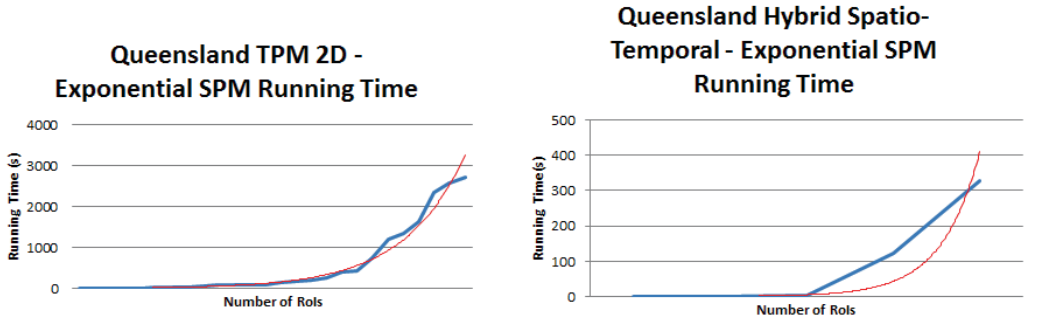

Figure 9: SPM running time.

5 Conclusion

In conclusion we have shown the merit in our spatio-temporal mining approach. We expected spatio-temporal mining to provide higher quality more useful patterns over traditional approaches. Results of our Queensland Flickr dataset have shown this is the case. The spatio-temporal approach uncovered interesting seasonal patterns along the east coast, and local yearly patterns in Brisbane. Such patterns are highly valuable to the tourism sciences industry, and thus this validates the usefulness of our approach. Overall, we acknowledge the effectiveness of our approach on Flickr data and suggest that our approach is generic enough to be applicable and useful to any field where moving entity trajectory data is abundant and collectable.

References

[1] Guochen Cai, Chihiro Hio, Luke Bermingham, Kyungmi Lee, and Ickjai Lee. Mining Frequent Trajectory Patterns and Regions-of-Interest from Flickr Photos. *47th Hawaii International Conference on System Sciences*, pages 1454–1463, January 6-9 2014.

[2] Guochen Cai, Chihiro Hio, Luke Bermingham, Kyungmi Lee, and Ickjai Lee. Sequential pattern mining of geo-tagged photos with an arbitrary regions-of-interest detection method. *Expert Systems with Applications*, 41:3514–3526, 2014.

[3] Fosca Giannotti, Mirco Nanni, Fabio Pinelli, and Dino Pedreschi. Trajectory pattern mining. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '07, pages 330–339, New York, NY, USA, 2007. ACM.

[4] Chihiro Hio, Guochen Cai, Luke Bermingham, Kyungmi Lee, and Ickjai Lee. A Hybrid Grid-based Method for Mining Arbitrary Regions-of-Interest from Trajectories. In *Proceedings of Workshop on Machine Learning for Sensory Data Analysis*, pages 5–12. ACM Press, 2014.

[5] Takeshi Kurashima, Tomoharu Iwata, Go Irie, and Ko Fujimura. Travel route recommendation using geotags in photo sharing sites. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, pages 579–588, New York, NY, USA, 2010. ACM.

[6] Xin Lu, Changhu Wang, Jiang-Ming Yang, Yanwei Pang, and Lei Zhang 0001. Photo2trip: generating travel routes from geo-tagged photos for trip planning. In Alberto Del Bimbo, Shih-Fu Chang, and Arnold W. M. Smeulders, editors, *ACM Multimedia*, pages 143–152. ACM, 2010.

[7] Kohya Okuyama and Keiji Yanai. A travel planning system based on travel trajectories extracted from a large number of geotagged photos on the web. In *The Era of Interactive Media*, pages 657–670. Springer New York, 2013.

[8] Jian Pei, Jiawei Han, Behzad Mortazavi-Asl, Helen Pinto, Qiming Chen, Umeshwar Dayal, and Meichun Hsu. Prefixspan: Mining sequential patterns by prefix-projected growth. In *Proceedings of the 17th International Conference on Data Engineering*, pages 215–224, Washington, DC, USA, 2001. IEEE Computer Society.

[9] Yue Shi, Pavel Serdyukov, Alan Hanjalic, and Martha Larson. Personalized landmark recommendation based on geotags from photo sharing sites. In Lada A. Adamic, Ricardo A. Baeza-Yates, and Scott Counts, editors, *ICWSM*. The AAAI Press, 2011.